

个性化信息流构建实践

陈伟平







1 产品背景



美柚: 平台型APP 700万+dau



POSSES. 10 15 3 投資收入 海田田 田田田 to a 医多线比较 1978 6,810 24 物的网络亚 THE TORN STOL COMMITTE 抗性性疾病 XX. **利尼拉公人**第766年6 48 英母親整治 自主和公共生和共 GREEN SULKNESS 兴集 在中部技术

记录: 经期记录, 工具功能

痛点: 使用周期短 需要才会来 用完马上走

愿景: 爱我别走 愿景: 提供有价值内容

社区: 有效提升活跃度及留存

痛点:数量少种类单一缺乏专业性



内容扩充



UGC:美柚社区

PGC:网易 腾讯 凤凰等专业媒体

OGC: papi酱 办公室小野 时尚芭莎等入驻自媒体平台

内容池:海量内容 数十万+信息过载

愿景:各取所需而非千人一面



如何把海量内容推给用户?

个性化信息流

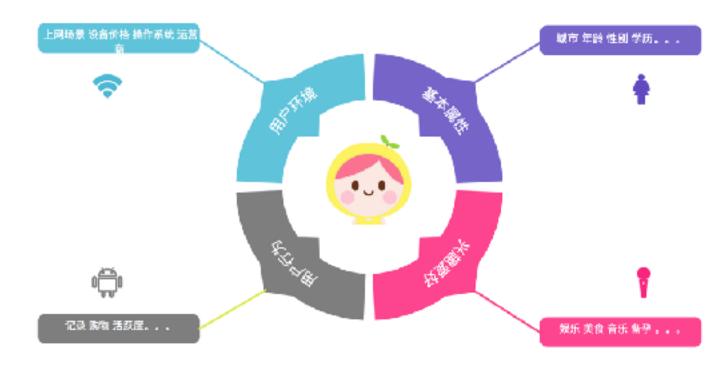




2 策略算法



推荐基础:用户画像







兴趣画像因素: Item画像(标签提取)

1基于K-means聚类提取

步骤: 取样本 分词 聚类 概括标签 计算词语标签向量 再计算文本标签

问题: 聚类时间长 数目不好控 难以概括每一类词语 新增困难

2 基于Bayes分类提取

问题:无法解决语料中一词多义和多词一义的问题 它更像是词法分析,而非语义分析

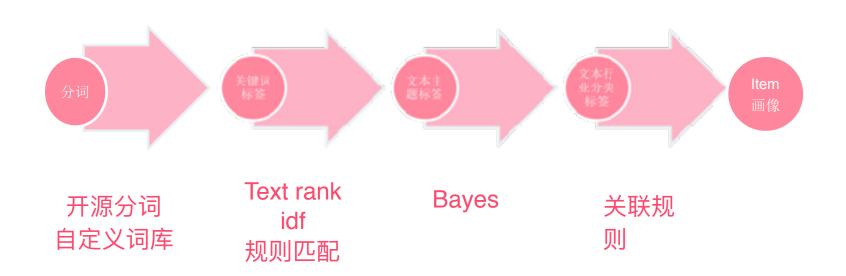
3 基于改进式Bayes分类提取



标签体系

行业分类 23 体育休闲 新闻时事 孕育指南 娱乐八卦 0 0 0 主题标签 74 星座 情感 明星 居家 0 0 0 关键词标签10万+ 宝宝 周杰伦 跑男 婆婆

文本内容识别步骤



兴趣画像: 行为上报 权重更新

用户行为

- 点击item
- 曝光item
- 回复
- 收藏。。

实时处理

- Kafka
- Sparkstreaming

用户画像

- 权重+(click+1, collect+2)
- 权重- (曝光 time)

策略算法













实时推荐





记录功能推荐 打通app内数据 实时I2I 相似推荐



冷启动

问题: 新用户没有行为

办法: 大数据分析 寻找共性

女性身份信息 设备信息 地域

- 0
- 0
- 0





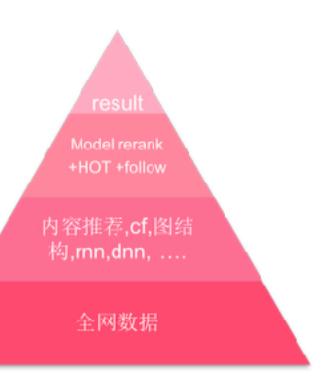




重排序

用户特征: 静态特征 动态特征

Item特征:静态特征(长度 类型 图片 数。。) 动态特征 (ctr,停留时长, 点赞,收藏。。。)



其他策略

运营要闻推荐 新内容定向投放 关注用户 活动专题 多样性 小说





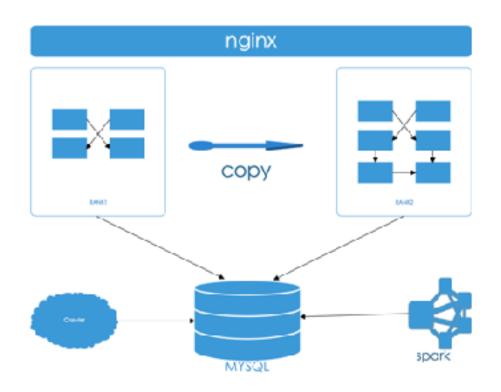


Start: 想什么做什么 算法单一

逻辑简单

灰度: 快速上线 copy

架构1.0: 快





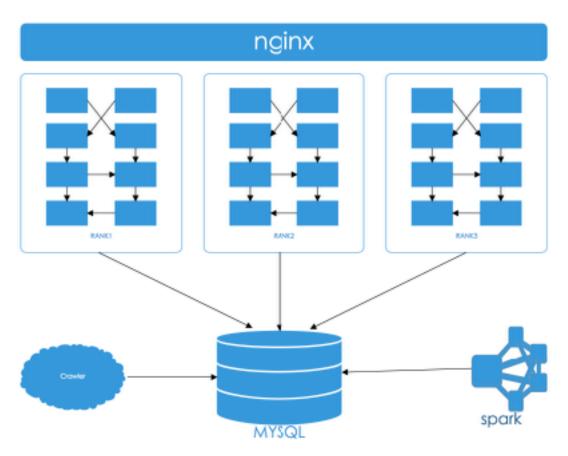
业务: 增长快 需求多 逻辑多

服务: 部署慢 分发乱

代码: 管理复杂 n个版本

问题: 难以维护 接近崩溃







架构1.0



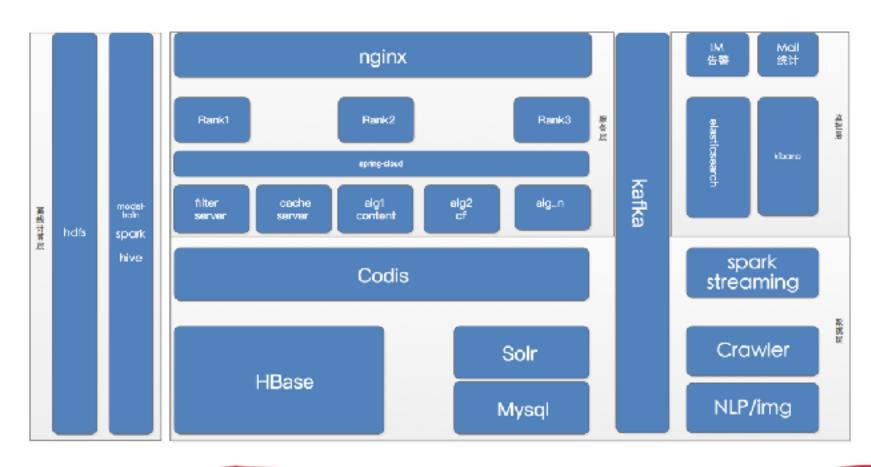


架构2.0



模块多 依赖大 环境复杂 难以升级 可插拔(模块解耦) 高可用 (容错监控) 灵活扩展 快速灰度







	架构1.0(单体服务)	架构1.1(共用库)	架构2.0(微服务)
易用性	本地调用(胜)	共享库引用 + 本地调用	API 调用
系统设计	高手能实现低耦合 数据最终还是耦合一起 一荣俱荣 一损俱损	在使用方的源代码中引用共享 库的类和方法 数据耦合	业务包装成服务,数据 代码物理上隔离,实现 高内聚低耦合相对容易 (胜)
升级,Bug修复	整体变动,容易导致不相 关模块的意外失败	改源代码,使用方重新打包	单独升级,其它模块无 感知,容易实施敏捷开 发方法(胜)
运维效率	简单直接(胜)	相对简单	大系统被拆分为多个小系统,部署和运维难度加大,但可以利用DevOps等方式将运维工作自动化



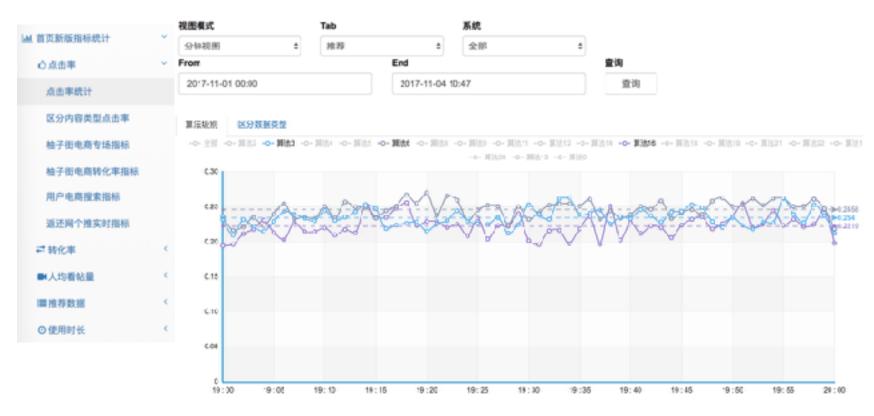
心得: 结合业务 快速迭代 适时重构







实时指标统计



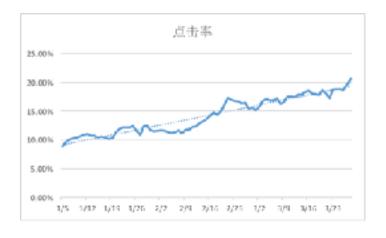
效果评估





用户轨迹查询 异常排查









Thank you

